修士論文要旨

Developing feasibility of the Chinese Learners' Speech Corpus referring to the Corpus of Spontaneous Japanese(CSJ)

Raymond Shen Director : Hideaki Kikuchi

1 Introduction

In the currently existing Chinese corpus, speech corpora are definitely the minority. Furthermore, there are few speech corpora identifying Chinese as an inter-language. Accordingly, the Chinese Learner's Speech Corpus shows the importance as well in the research of second language acquisition.

In this study, I am attempting to construct a pilot version of the Chinese Learner's Speech Corpus referring to CSJ and meanwhile, verify the developing feasibility of it.

Referring to the results of the questionnaires, when transcribing the collected speech data, I also annotated some information which can realize the evaluation of the pilot version. Then I focused on the XML documentation, which is a feature of CSJ and proved effective in eliminating conflicts occurring among various kinds of additive information on one and the same speech in CSJ. The transcription data with morphological and label information integrated will be transformed into the XML documents and through which grammatical and morphological phenomena can be retrieved conveniently.

2 Previous Studies

I mainly referred to the tentative ideas of constructing Chinese Learners' Spoken Corpus. And the construction of CSJ, especially the part explaining XML document. From the former study, I confirmed the feasibility of constructing a pilot version of Chinese Learners' Speech Corpus by collecting a few data and however, from the latter, I hypothesized that the XML format of data, which has been proved effective in CSJ, could also be used in my pilot version. In CSJ, due to the various kinds of additive information, conflicts occur and the testifying of consistency is sabotaged.

I thought Chinese Learners' Speech Corpus can be constructed by referring to the XML documentation part in CSJ. To prove the feasibility, I decided to obtain some advice and opinions from the assuming users.

3 Questionnaires

I have received eight questionnaires from teachers of teaching Chinese as a foreign language with different backgrounds in teaching years and experiences respectively.

Regarding of the opinions given by the experienced teachers, I consider them valuable and become to focus on four tones, word order, interfering from mother tongues and etc. I also got theoretical basis for the quantity of data required, annotations, functions, kinds of information with which I can execute the construction into details.

4 Details of corpus

4.1 Data sources and processing

I collected one speech of about 9 speakers each, approximately 100 minutes totally and 3-4 native languages are involved. The study levels of speakers vary from the novice to the superior. The lengths of speeches are about 10-25 minutes, which are considered enough for grammatical or morphological analysis according to the results of questionnaires survey mentioned above. Data are totally kept anonymous and collected by desktop microphone or telephone.

In data processing, first I picked out proper data and then divide each into utterances. In the pilot version, considering about the emphasis are put on grammatical disfluency and morphological information, I divided utterances into segments with complete meanings and write it clearly as a dividing rule in the data processing manual. Furthermore, I also transcribe the utterances and annotate mistakes and disfluency occurred in the speech with the results saved in files respectively.

4.2 XML documentation

Considering features of the inter-language corpus, I built a hierarchical data structure of the base XML. On the top there is the TALK element, which is to be derived from the information about each speech data including speaker's information. The next level is U (utterance) element, who has two children, WU (word unit) and GD(grammatical disfluency). The information of U element is mainly from transcriptions. WU element contains pronunciation and morphological analysis, however, GD element represents grammatical disfluency occurred in the speech data. At bottom of the hierarchy, the D (disfluency) element, whose parent is the WU element, consists of all the disfluency (not grammatically), such like the mispronunciation in four tones, muttering, repeating and so on.

4.3 Annotation

To prove the feasibility and efficiency of the pilot version I constructed, evaluations are certainly required.According to the questionnaire survey mentioned above, the pilot version of Chinese Learners' Speech Corpus can be evaluated in various ways.

For example, the situations of disorder can be understood at one glance by searching the tag<O>, which is annotated on sentences in wrong order. The original texts can also be called out and the proportion of disorder can be given if further views are necessary. Moreover, as a big problem in Chinese learning, the use of particle such like '的' and '了' seems to be paid attention to by both teachers and learners. In this case, I annotated the misuse and neglect of particle as tag<P> so that users can grasp the situation by searching the tag. Besides, there are a few other items like above to evaluate this pilot version. Thanks to the XML documents, mistakes and disfluency occurred in the speech data can be easily retrieved with time information and original texts as well by using the XML browser.

5 Conclusion and Future Work

The Chinese Learner's Speech Corpus can function as an instruction for both teachers and learners, and be used as a reference of compiling teaching materials, too.

The data collected are also far from adequate and various in languages for constructing a full version of Chinese Learners' Speech Corpus.